The Wikification of the Brede Database

Finn Årup Nielsen

Lundbeck Foundation Center for Integrated Molecular Brain Imaging
at
DTU Informatics and Neurobiology Research Unit, Copenhagen
University Hospital Rigshospitalet

April 1, 2008
Abstract

Wikipedia is presently among the top-ten visited web-sites and there are now over 10 million articles across the different language editions. Templates and categories in Wikipedia allow the editors to put structure on the information in Wikipedia. This kind of content can be extracted from Wikipedia and analyzed offline. Structured information may also be added automatically to Wikipedia from other web-based databases. Furthermore, specialized wikis inspired by the idea of the semantic web forms the so-called semantic wikis, where information is kept structured. I will show examples of my analysis of scientific information on Wikipedia’s structured content and show some of my initial efforts in adding wiki-functionality to the Brede Database — or Brede functionality to Wikipedia.
The wiki-idea with Internet-based shared and easy editing started by Ward Cunningham in the middle of the nineties (Leuf and Cunningham, 2001).

Wikipedia started in 2001 and experienced exponential growth in the beginning.

According to Alexa Wikipedia is presently on the top-ten over most visited web-sites.

Figure 1: Growth in the number of articles on the English Wikipedia. From the Wikipedia article *Wikipedia:Modelling Wikipedia’s growth.*
Structured content in Wikipedia

The content in Wikipedia may contain structured information, e.g., scientific citations:

```
{{cite journal|author=Filipek PA, Accardo PJ, Baranek GT ’’et al.’’
|title=The screening and diagnosis of autistic spectrum disorders
|journal=J Autism Dev Disord |date=1999 |volume=29 |issue=6
|pages=439--84
|doi=10.1023/A:1021943802493|pmid=10638459}}
```

Or Proteins:

```
{{protein
|Name=5-hydroxytryptamine (serotonin) receptor 2A
|caption= |image= |width= |HGNCid=5293 |Symbol=HTR2A
|AltSymbols=HTR2 |EntrezGene=3356 |OMIM=182135 |RefSeq=NM_000621
|UniProt=P28223 |PDB= |ECnumber= |Chromosome=13 |Arm=q |Band=14
|LocusSupplementaryData=-q21}}
```
Year of publication of cited journal article

Wikipedia output journal citations

Number of citations

Year


0 500 1000 1500 2000 2500 3000 3500 4000 4500 5000 5500 6000
Analysis of scientific citations

Scientific citations from Wikipedia to journals

Comparison against Journal Citation Reports from Thomson Scientific

The *Nature* and *Science* journals are the most cited from Wikipedia.

The product between total citations and impact factor showed good correlation with the Wikipedia citation count (Nielsen, 2007).
Wikification of the Brede Database

- JCR total citations x impact factor
- Wikipedia citations

- Science
- Nature
- PNAS
- NeuroImage
- NeuroReport
- Archives of General Psychiatry
- Biological Psychiatry
- Neuroscience Letters
- European Journal of Neuroscience
- Neuroscience
- JNeurosci
- NatMed
- AnnNeurol
- Neurology
- Neuroscience
Clustering of Wikipedia data

Clustering of citations from Wikipedia articles to scientific journals from a October 2007 dump.

Data matrix (Wikipedia article × journal)

Non-negative matrix factorization with increasing number of clusters (components) (Nielsen et al., 2005)

Clusters related to astrophysics, medicine, intelligens, immunology, bacteria, …

<table>
<thead>
<tr>
<th>Cluster</th>
<th>List of molecules</th>
<th>List of scientific</th>
<th>Myocardial infarction</th>
<th>History of invasive cancer</th>
<th>Obesity</th>
<th>Patient safety</th>
<th>Race and intelligence</th>
<th>IQ and the Wealth Gap</th>
<th>Sex and intelligence</th>
<th>Flynn effect</th>
<th>Henry George Fo</th>
<th>IQ and the Wealth Gap</th>
<th>Sex and intelligence</th>
<th>Flynn effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>List of molecules</td>
<td>List of scientific</td>
<td>Myocardial infarction</td>
<td>History of invasive cancer</td>
<td>Obesity</td>
<td>Patient safety</td>
<td>Race and intelligence</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
<td>Henry George Fo</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
</tr>
<tr>
<td>2</td>
<td>List of molecules</td>
<td>List of scientific</td>
<td>Myocardial infarction</td>
<td>History of invasive cancer</td>
<td>Obesity</td>
<td>Patient safety</td>
<td>Race and intelligence</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
<td>Henry George Fo</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
</tr>
<tr>
<td>3</td>
<td>List of molecules</td>
<td>List of scientific</td>
<td>Myocardial infarction</td>
<td>History of invasive cancer</td>
<td>Obesity</td>
<td>Patient safety</td>
<td>Race and intelligence</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
<td>Henry George Fo</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
</tr>
<tr>
<td>4</td>
<td>List of molecules</td>
<td>List of scientific</td>
<td>Myocardial infarction</td>
<td>History of invasive cancer</td>
<td>Obesity</td>
<td>Patient safety</td>
<td>Race and intelligence</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
<td>Henry George Fo</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
</tr>
<tr>
<td>5</td>
<td>List of molecules</td>
<td>List of scientific</td>
<td>Myocardial infarction</td>
<td>History of invasive cancer</td>
<td>Obesity</td>
<td>Patient safety</td>
<td>Race and intelligence</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
<td>Henry George Fo</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
</tr>
<tr>
<td>6</td>
<td>List of molecules</td>
<td>List of scientific</td>
<td>Myocardial infarction</td>
<td>History of invasive cancer</td>
<td>Obesity</td>
<td>Patient safety</td>
<td>Race and intelligence</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
<td>Henry George Fo</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
</tr>
<tr>
<td>7</td>
<td>List of molecules</td>
<td>List of scientific</td>
<td>Myocardial infarction</td>
<td>History of invasive cancer</td>
<td>Obesity</td>
<td>Patient safety</td>
<td>Race and intelligence</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
<td>Henry George Fo</td>
<td>IQ and the Wealth Gap</td>
<td>Sex and intelligence</td>
<td>Flynn effect</td>
</tr>
</tbody>
</table>
Growth in citations from Wikipedia . . .

The use of the cite journal template has increased since 2006.

Over the new year 2007/2008 a bot adds numerous pages for proteins with a lot of scientific citations expanding the number of citations from 74,776 citations in the October 2007 dump to 228,593 in the March 2008 dump.
## Most cited journals in March 2008

<table>
<thead>
<tr>
<th>Citations</th>
<th>Journal name</th>
</tr>
</thead>
<tbody>
<tr>
<td>16739</td>
<td>The Journal of Biological Chemistry</td>
</tr>
<tr>
<td>12779</td>
<td>PNAS</td>
</tr>
<tr>
<td>8772</td>
<td>Genome Research</td>
</tr>
<tr>
<td>7561</td>
<td>Nature</td>
</tr>
<tr>
<td>4007</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>3928</td>
<td>Genomics</td>
</tr>
<tr>
<td>3689</td>
<td>Science</td>
</tr>
<tr>
<td>3511</td>
<td>Gene</td>
</tr>
<tr>
<td>3380</td>
<td>Biochemical and Biophysical Research Communications</td>
</tr>
<tr>
<td>3043</td>
<td>Molecular and Cellular Biology</td>
</tr>
<tr>
<td>2975</td>
<td>Cell</td>
</tr>
<tr>
<td>2261</td>
<td>The EMBO Journal</td>
</tr>
</tbody>
</table>

Table 1: Most cited journals from Wikipedia in the 12th March 2008 dump.
## Clusters in the new dump

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Wikipedia hub articles</th>
<th>Authoritative journals</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Cancer’</td>
<td>RBL2, MYB, ERG (gene), EPS8</td>
<td>Oncogene, Cancer Research, Int. J. Cancer, Gene &amp; Development</td>
</tr>
<tr>
<td>‘Immunology’</td>
<td>DNA vaccination, CCL21, HLA-DQ8, HLA-DQA1</td>
<td>The Journal of Immunology, The Journal of Experimental Medicine, Tissue Antigens, Eur. J. Immunol.</td>
</tr>
</tbody>
</table>

Table 2: The top Wikipedia hubs articles and authoritative journals with respect to clusters from a non-negative matrix factorization with twenty clusters.
Other type of analysis. Preliminaries

Construct binary matrix $X(\text{articles} \times \text{authors})$ with 1 indicated an edit.

Excluding usernames matching “bot” and documents beginning with “Wikipedia”.

Exclude articles with less than three different authors.

Danish Wikipedia: $X(12774 \times 3149)$ with density 0.0025

Some kind of normalization? The results may depend on the exact kind.

Again non-negative matrix factorization (Lee and Seung, 2001) — one of the algorithms “off the shelf” in the Brede Toolbox (Nielsen and Hansen, 2000).
Wikipedia clustering. Some cluster example

Danish Kings: Christian 3., Christoffer 1., Erik Klipping, Frederik 1.

Countries: Portugal, Slovenien, Polen, Tyskland, Belgien, Estland


Danish munipalities and counties: Roskilde Amt, Birkerød Kommune, Frederikssund Kommune


Discussion: Bruger diskussion:User#1, Bruger diskussion:User#2, Jesus fra Nazaret, Kristendom, Anders Fogh Rasmussen, Diskussion:Dansk Folkeparti, Diskussion:Muhammed-tegningerne, Kreationisme
There exist tools that helps with the construction of Wikipedia templates.

For scientific journal articles in Wikipedia:

Zotero reference manager, a plugin to the Firefox web-browser

“Wikipedia template filling”, a web-service
Construction with bots

Protein Box Bot builds pages with proteins and may update whole or part of the page based on other databases. It builds an "Infobox", abstracts and citation information.

“Assisted editing”/validation
(Iorio and Zacchirolı, 2006)

“Can we link it”

Figure 2: Wikipedia article for gene where all information has been entered via a bot.
Large-scale databasing Wikipedia templates

DBpedia: Extract templates from Wikipedia and put them in a database (Auer and Lehmann, 2007; Auer et al., 2008)

Basic datatype is a triple.

DPpedia is presented on the web with search interface: SPARQL, “Leipzig query builder”
Scientific and semantic wiki

Example question: “What are the hundred world-largest cities with a female mayor?”

OmegaWiki, Semantic Mediawiki (Völkel et al., 2006), WikiFactory, OntoWiki (Auer et al., 2006), WikiProteins
Neuroimaging data on Wikipedia?

Series of values may be re-present in a table: Here in-formation from neuroimag-ing DASB studies.

While this representation may give an overview that can be updated incrementally by different individuals, it may not be a good representation to work with wrt. machine reading.
The Brede Database presently contains information about functional neuroimaging using XML files for storage of data with entry function implemented in Matlab and some search facility implemented in Perl.

Figure 3: Screenshot of a program to entry of data.
<Bibs>
  <Bib>
    <wobib>1</wobib>
    <type>bib</type>
    <Author>
      <surname>Law</surname>
      <firstname>Ian</firstname>
      <name>Ian Law</name>
    </Author>
    ...<citation>Brain 1998 Nov;121 ( Pt 11):2189-200.</citation>
    ...
    <title>Parieto-occipital cortex activation during self-generated eye movements in the dark</title>
    ...
    <Exp>
      <capsuleDescription>Self-generated saccades</capsuleDescription>
      <freeFormDescription>Voluntary self-generated large-amplitude horizontal saccades with eyes open versus resting state with eyes closed</freeFormDescription>
      ...
    <Loc>
      <functionalArea>Left frontal eye field</functionalArea>
      <brodmann></brodmann>
      <zScore>4.82</zScore>
      ...
    </Loc>
  </Bib>
</Bibs>
Wikification of the Brede Database

Extracting data from Wikipedia to matlab

With the Brede Toolbox extract cited journals from the *Altanserin* article:

```
s = brede_web_wikipedia('Altanserin');
S = brede_str_wptemplate2struct(s,'template','Cite journal');
B = brede_bib_bib2bib(S);
brede_ui_bibs(B)
```

This will bring up an Matlab GUI with bibliographic information about all the cited journals.

There are obstacles for extraction of, e.g., nested templates (templates within templates) and fields in templates with multiple entries (e.g., the author field in the cite journal template)
Representation of Talairach coordinates

Template arguments with integer parameters:

```latex
{{Talairach coordinates
 | x_1=-38 | y_1=2 | z_1=37
 | x_2= 48 | y_2=-40 | z_2 =9 | anatomy_2 =Right temporoparietal junction
 | x_3=51 | y_3=15 | z_3=34 | functional_area_3 =[[Supplementary motor area]] }}
```

Or with a template within a template:

```latex
{{Talairach coordinates
 | Coordinates =
   {{Talairach coordinate | x=-38 | y=2 | z=37}}
   {{Talairach coordinate | x= 48 | y=-40 | z=9 | anatomy=Right temporoparietal junction }}
   {{Talairach coordinate | x=51 | y=15 | z=34
    | functional_area=[[Supplementary motor area]] }} }}
```
Talairach coordinates

Right Temporoparietal Cortex Activation during Visuo-proproceptive Conflict is a scientific article describing an functional magnetic resonance imaging (fMRI) experiment. It was published in 2004:


The table with Talairach coordinates is constructed from a nested template.

<table>
<thead>
<tr>
<th>#</th>
<th>Anatomy</th>
<th>Functional area</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Left precentral gyrus/frontal lobe</td>
<td></td>
<td>-38</td>
<td>2</td>
<td>37</td>
</tr>
<tr>
<td>2</td>
<td>Right temporoparietal junction</td>
<td></td>
<td>48</td>
<td>-40</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>Right precentral gyrus/frontal lobe</td>
<td>Supplementary motor area</td>
<td>51</td>
<td>15</td>
<td>34</td>
</tr>
</tbody>
</table>

Figure 4: Formatted Talairach coordinates in Wikipedia.
Extensible database for Brede

Extensible database: Add new fields ("columns")

Basic datatype a Quintuple: structure, field name, value, rank, link.

Will this be useful for editing and display?
Conclusion

Wiki and Wikipedia content may not only be source for “human” reading and editing but also:

Automatic adding information

Extraction of information

A source for data that may be subjected to computer-based analysis
References


Auer, S., Riechert, T., and Dietzold, S. (2006). OntoWiki — a tool for social, semantic collaboration. In Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Uschold, P., and Aroyo, L., editors, The Semantic Web — ISWC 2006, volume 4273 of Lecture Notes in Computer Science, Heidelberg/Berlin. Springer. Description of a collaborative wiki-based knowledge base where instances can be linked to classes and have properties. These properties can be, e.g., a date that may be displayed in a calendar or geographical coordinates that can be integrated with Google Map API. Users can edit the individual properties, helped by AJAX technology. The system also has versioning and some search facility. OntoWiki uses the Powl platform. This OntoWiki should not be confused with the Innsbruck OntoWiki.


The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/ for more information. Each version is given a distinguishing version number. If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and/or additional documents that qualify as Secondary Cover Texts, this License, specified as the “Licence”, applies to the Document with the modification that there must be included, immediately after the title page, an additional front-matter section若as “History” that explains how the Document was produced. The requirement to provide this section applies to the modified version as well as to the original version. It is requested, but not required, that you contact the authors of the Document or their publisher or copyright holder before redistributing modified versions. Instead of requiring all users to be able to freely copy your works, the Free Software Foundation, through the GNU Free Documentation License, grants you permission to distribute copies to the public by distribution of printed or electronic copies of this license, as per the above requirements.地方以获得免费复制你的作品，Free Software Foundation，通过GNU Free Documentation License，授予你以按照上述要求将该许可传播给公众的权利。